Objective Speech Quality Estimates for Project 25/Voice over Long Term Evolution (P25/VoLTE) Interconnections

Public Safety Communications Technical Report

May 2013

Homeland Security

Science and Technology

Initial Release: March 2013

Public Safety Communications Technical Report

Objective Speech Quality Estimates for Project 25/Voice over Long Term Evolution (P25/VoLTE) Interconnections

DHS-TR-PSC-13-01
March 2013

Reported for: **The Office for Interoperability and Compatibility by the Public Safety Communications Research Program**

Homeland Security
Science and Technology

Intentionally Blank

Publication Notice

Disclaimer

The U.S. Department of Homeland Security Science and Technology Directorate (S&T) serves as the primary research and development arm of the Department, using our nation's scientific and technological resources to provide local, tribal, state, and federal officials with the technology and capabilities to protect the homeland. Managed by S&T's First Responders Group, the Office for Interoperability and Compatibility currently assists in the coordination of interoperability efforts across the nation.

Certain commercial equipment, materials, and software are identified to specify technical aspects of the reported procedures and results. In no case does such identification imply recommendations or endorsement by the U.S. Government, its departments, or its agencies, nor does it imply that the equipment, materials, and software identified are the best available for this purpose.

Abstract

This document contains requirements for an interoperable public safety broadband communications nationwide network to serve all local, tribal, state, and federal first responder communications.

Contact Information

Please send comments or questions to: SandTFRG@hq.dhs.gov

Intentionally Blank

Contents

Intentionally Blank

Abstract

In an extended Project 25/Voice over Long Term Evolution (P25/VoLTE) public safety communication system, voice signals will pass through both Multi-Band Excitation (MBE) and Adaptive Multi-Rate (AMR) speech coders. Thus, it is important to quantify the speech quality that can be expected for MBE, AMR, and combinations of these speech coders. We used the Perceptual Evaluation of Speech Quality (PESQ) algorithm to provide initial assessments of speech quality for these coders alone and in combinations.

Funding for this work was provided by the U.S. Department of Homeland Security's (DHS) Science and Technology Directorate, through its Office for Interoperability and Compatibility. The work was conducted at the Public Safety Communications Research program's Boulder, Colorado, laboratories under the supervision of J. R. Bratcher. Certain commercial equipment and materials are identified in this report to adequately specify the technical aspects of the reported results. In no case does such identification imply recommendation or endorsement by the National Telecommunications and Information Administration, National Institute of Standards and Technology, or DHS, nor does it imply that the material or equipment identified is the best available for this purpose.

Index Terms: AMBE, AMBEoLTE, PESQ, P25, Speech Quality, VoLTE

1 Introduction

The P25 digital radio system is widely used for public safety voice communications in the U.S.; however, due to data rate limitations, a low bit rate speech coder is required. The solution adopted in P25 is MBE speech coding. It is anticipated that the broadband public safety network to be built by FirstNet will employ LTE radio technology, and that voice communications will be accomplished through VoLTE. Because LTE is a broadband network, higher rate speech coding is allowed and a popular choice for VoLTE is the AMR speech coder. Communications between P25 systems and VoLTE systems will be required. In these cases, speech signals will pass through both MBE and AMR speech coders. Each speech coder adds some distortion to the speech signal and reduces the resulting speech quality. Therefore, it is important to quantify the speech quality that can be expected when MBE and AMR speech coding are combined and to compare these quality levels with those associated with MBE and AMR in isolation.

Speech quality is most directly assessed through subjective testing. In these tests, subjects listen to recordings in a controlled laboratory environment and rate the speech quality of each recording on a numerical scale. A popular scale is the mean opinion score (MOS) scale where the numbers 1, 2, 3, 4, and 5 are associated with the speech quality levels bad, poor, fair, good, and excellent, respectively.

Speech intelligibility can also be tested subjectively. One protocol is the Modified Rhyme Test (MRT). Subjects in an MRT hear a word and have to select that word from a set of similar sounding words (e.g., wed, fed, shed, red, bed, led) on a scoring form. After many repetitions using different sets of words, the percentage of correctly identified words provides a measure of intelligibility for the system under test. Subjective tests, be they speech quality tests or speech intelligibility tests, require significant time, effort, and specialized equipment.

There is a less costly alternative to subjective testing that can provide very good initial indications of speech quality. The Perceptual Evaluation of Speech Quality (PESQ) algorithm has been standardized by the International Telecommunications Union, Telecommunication Standardization Sector (ITU-T) as Recommendation P.862. This algorithm can process digital speech recordings in real time and provide estimates of the speech quality associated with those recordings. We chose to use PESQ to provide initial assessments of the speech quality associated with various combinations of MBE and AMR speech coding.

This report is organized as follows:

- In Section 3, we briefly describe MBE speech coding, AMR speech coding, and the PESQ algorithm.

- In Section 4, we explain how we use these tools and speech recordings to evaluate the speech quality for 42 different scenarios.
- Finally, in Section 5 we present and discuss the results of this evaluation.

D.J. Atkinson of the Public Safety Communications Research (PSCR) program conducted the initial work in this area in December 2011, and he presented those results at the International Wireless Communications Expo on February 22, 2012. Further work was conducted in July 2012, and this report was produced to provide a detailed archival record of the entire effort.

2 Speech Coders and Quality Estimation

Digital speech coders reduce the bit rate required to transmit digital speech while preserving speech quality to the extent possible. In general, however, reducing the bit rate requires some sacrifice of speech quality. Several additional factors must be considered when designing or selecting a speech coder for a given application. These factors include complexity, delay, robustness, and flexibility.

The MBE family of speech coders [1], [2], and [3] used in P25 offers relatively low bit rates and relatively high robustness to acoustic background noises that may be competing with the desired speech signal at the transmit location. These desirable traits are achieved by analyzing 20 millisecond (ms) frames of an incoming speech signal in multiple spectral bands to determine the mixture of periodic and non-periodic signal in each band. Our work used the Advanced MBE+2TM (AMBE+2) speech coder as implemented in a Windows Command Line Interface (CLI) executable provided by Digital Voice Systems, Inc. (DVSI). This executable was further labeled as "floating point version 1.40e, October 14, 2009." This version of the DVSI research and development DOS executable code is equivalent to the digital signal processing (DSP) production code version 1.6 that is currently offered in P25 radios. We used both full rate (7200 bits per second [(bps)]) and half-rate (3600 bps) modes.

The narrowband AMR speech coder is a popular choice for cellular phones [4] and [5]. It offers speech coding at eight different bit rates (i.e., 4.75, 5.15, 5.9, 6.7, 7.4, 7.95, 10.2, and 12.2 kbps) and can change its rate at the start of any 20 ms speech frame. When radio conditions deteriorate, the speech coding rate can be reduced, thus allowing for the addition of more robust channel coding. The goal is to allow a conversation to continue, albeit with lowered speech quality. This is a desirable alternative to unintelligible speech or a dropped call. In addition, AMR features a voice activity detector, a comfort noise generator, and an error concealment mechanism. AMR speech coding is successfully used in third-generation (3G) wireless voice services.

The underlying core speech coding algorithm in AMR is called algebraic code-excited linear prediction (ACELP). Linear prediction provides a popular and efficient speech representation consisting of a linear filter and an excitation signal. An ACELP encoder determines the proper linear filter and excitation signal for a group of speech samples. Filter parameters are transformed and quantized for efficient transmission to the decoder. The encoder represents the excitation signal as a combination of vectors already stored in common "codebooks" at the encoder and decoder. The encoder can thus transmit to the decoder just the pointers to the proper codebook locations, rather than the actual vectors or the excitation signal itself. This allows for a great savings in transmitted data. The decoder can then reconstruct an approximation of the speech samples by passing the reconstructed excitation signal through the filter. We used the floating point Windows CLI executable AMR implementation provided by VoiceAge Corporation through the Open AMR Initiative. This software is compliant with 3GPP TS 26.071. For completeness, we investigated each of the eight available AMR bit rates.

The PESQ algorithm [6], [7], and [8] provides estimates of speech quality by comparing two digital speech files. One file contains the speech input to the system under test and the other contains the output from the system under test. The system under test could be one or more speech coders and decoders or other radio or telecommunications equipment. In order to emulate human hearing, PESQ transforms those signals into a perceptual domain. Key elements of this transformation are a non-linear frequency scale and a compressed intensity scale. PESQ then compares the transformed signals in a way that emulates human judgment. If this comparison stage finds a small perceptual difference between the two files, a high speech quality estimate is returned. If a larger difference is found, then a lower speech quality estimate is returned.

The main PESQ output is on a scale analogous to the MOS scale described in Section 2. It is called "MOS Listening Quality, Objective" (MOS-LQO). Like the original MOS scale, MOS-LQO runs from 1 (bad) to 5 (excellent). PESQ has been studied extensively, and MOS-LQO values are often found to be excellent surrogates for actual MOS values. The PESQ algorithm we used is a standards-compliant DOS executable implementation provided by OPTICOM GmbH.

3 Test Method

Our goal was to find MOS-LQO values for various speech coder configurations. Speech coder performance varies with talker traits and speech content. To take this variation into account, we used a group of 64 speech recordings with speech from four female talkers and four male talkers. There are eight recordings for each talker. Each recording contains two sentences from the Harvard phonetically balanced sentence lists [9]. Because the talkers were recorded in a sound isolation chamber, the recordings contain virtually no background noise. No background noise was added to the recordings. Altogether these 64 recordings contain about 7 minutes of speech in American English. We adjusted the level of each of the recordings to the correct level for the software speech coders. (This level is called "28 dB below clipping," which refers to the relationship between the Root Mean Square (RMS) speech signal level and the level at which the digital representation used in the speech coders would saturate.)

The generic block diagram for our speech processing work is given in Figure 1. First, speech is passed through an initial speech encoder and decoder (once-coded speech), connected by a perfect channel (meaning that the decoder receives the exact bitstream produced by the encoder). The decoded speech is then optionally passed through a second encoder and decoder (twice-coded speech), also connected by a perfect channel. Bit errors and lost packets are outside the scope of this investigation. Finally, the PESQ algorithm processes the original speech file and the final output speech file, comparing them to produce a value of MOS-LQO, as described in Section 3.

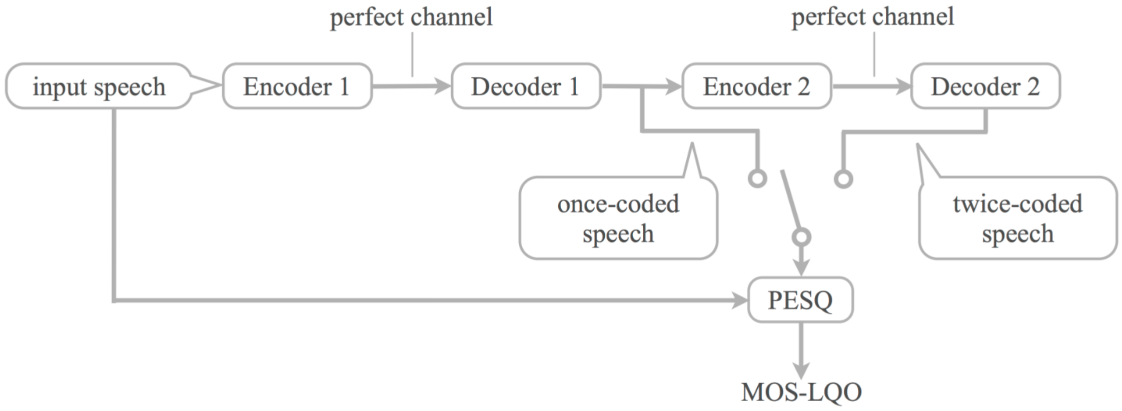


Figure 1: Block diagram for speech processing path resulting in MOS-LQO (estimated speech quality) values. For single speech coder scenarios, switch is in left position and once-coded speech is analyzed by PESQ. For interconnection scenarios, switch is in right position and twice-coded speech is analyzed by PESQ.

For completeness, we considered all eight available AMR bit rates and the two available MBE bit rates. In total, 10 single speech coder scenarios (once-coded speech) provided baselines against which we can compare the interconnection scenarios (twice-coded speech). For interconnection scenarios, we considered both the P25 to VoLTE scenario (MBE first, then AMR) as well as the VoLTE to P25 scenario (AMR first, then MBE). This produced (2 × 8) + (8 × 2) = 32 interconnection scenarios and a grand total of 10 + 32 = 42 scenarios. For each scenario, all 64 speech files are processed as shown in Figure 1, and 64 MOS-LQO values are produced.

4 Results and Discussion

For each scenario, the sample mean of the 64 MOS-LQO values gives the average estimated speech quality for that scenario. Associated with each sample mean is a 95 percent confidence interval, indicating the range in which the true mean must be, with 95 percent confidence. These means and confidence intervals are shown for the 10 single-speech coder scenarios in Figure 2. We display the results as a function of the speech coder bit rate. The rates commonly used to describe half- and full-rate MBE are 3600 and 7200 bps, respectively, but these rates include forward error correction and the actual speech coding rates are 2450 and 4400 bps, respectively. The relationship between estimated speech quality with no background noise and bit rate is clear. We can expect that when no background noise is present, VoLTE can provide speech quality equivalent to or better than P25, even when LTE radio conditions are so severe that the AMR coder is forced to operate at its lowest rate.

Means and 95 percent confidence intervals for the 16 interconnection scenarios involving full-rate MBE are shown in Figure 3a. Results for the P25 to VoLTE scenario (MBE first, then AMR) are shown in dark gray. Results for the VoLTE to P25 scenario (AMR first, then MBE) are shown in light gray. The speech quality differences between these two scenarios are not statistically significant at the 95 percent level.

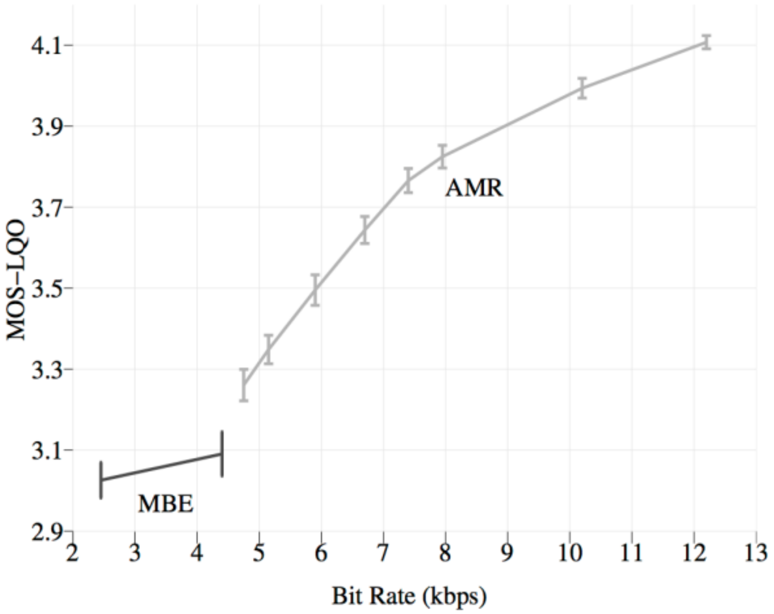


Figure 2: MOS-LQO means and 95 percent confidence intervals for MBE (two rates) and AMR (8 rates).

For comparison, the figure also shows the mean MOS-LQO for full-rate MBE alone as a back line across the top of the figure. The 95 percent confidence interval for that mean is shown as a shaded region about that line, extending across the top of the figure. When VoLTE and P25 are interconnected, the speech coding distortions accumulate, and the overall quality can be no better than the quality achieved by P25 alone. Therefore, Figure 3a shows MOS-LQO dropping from the reference level associated with MBE. The 95 percent confidence intervals provide useful visual indications of the uncertainty associated with each of the calculated means.

Interconnecting VoLTE with P25 reduces mean MOS-LQO values from the levels produced by P25 alone. But in which cases is that reduction a statistically significant one? To answer this question, we employed a two-tailed *t*-test with a 95 percent significance level. The null hypothesis for this test is "AMR/full-rate MBE combinations have the same MOS-LQO as full-rate MBE alone." For all 16 scenarios, the calculated *t* values exceed the threshold, and this null hypothesis should be rejected. In other words, in all 16 scenarios, the difference between the two is statistically significant.

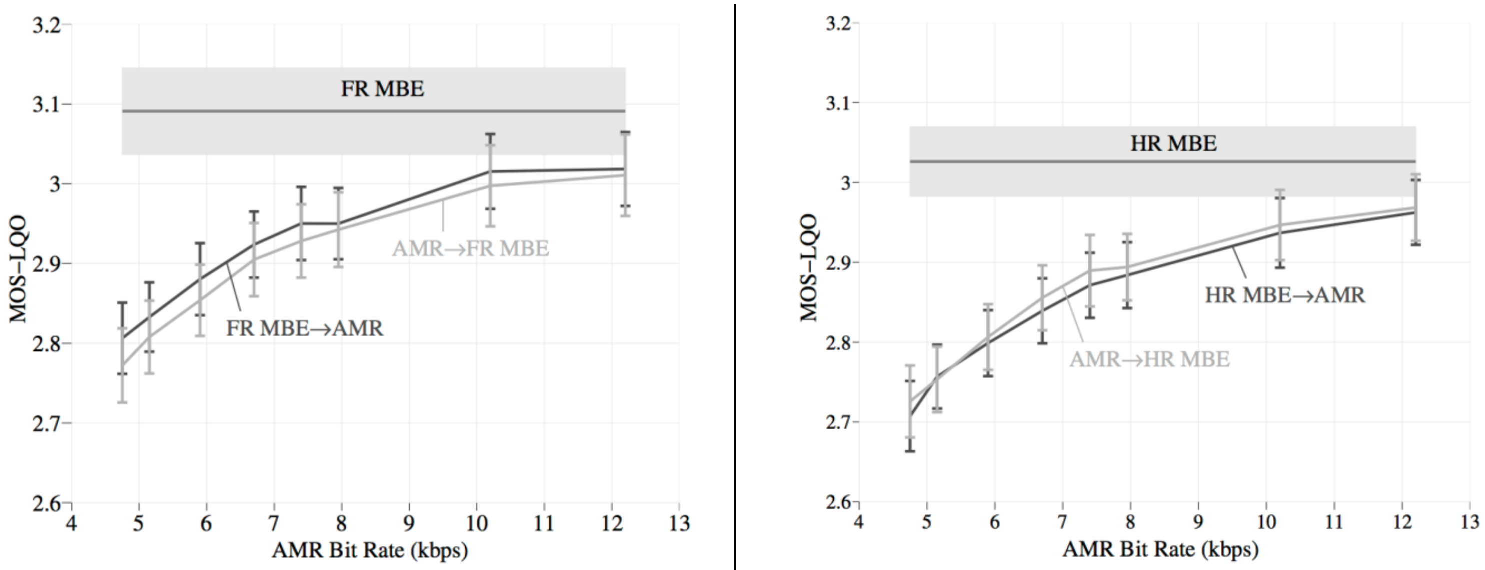
Figure 3b is analogous to Figure 3a, but half-rate MBE replaces the full-rate version. As in Figure 3a, the results for the P25 to VoLTE scenario (MBE first, then AMR) are shown in dark gray and those for the VoLTE to P25 scenario (AMR first, then MBE) are shown in light gray. The speech quality differences between these two scenarios are not statistically significant at the 95 percent level. The half-rate MBE result (mean and 95 percent confidence interval) is shown across the top of the figure. Again, Figure 3b shows MOS-LQO dropping from the reference level associated with half-rate MBE alone.

We again employed a two-tailed *t*-test with a 95 percent significance level to determine which of the AMR/MBE combinations in Figure 3b has an MOS-LQO that is significantly lower that that of MBE alone. The null hypothesis for this test is "AMR/half-rate MBE combinations have the same MOS-LQO as half-rate MBE alone." The calculated *t* values indicate that in all scenarios, except for the scenario of AMR coding at 12.2 kbps followed by half-rate MBE coding (far right, light gray), the null hypothesis should be rejected. In other words, in 15 of the 16 scenarios, the difference between the two is statistically significant. But AMR coding at 12.2 kbps followed by half-rate MBE coding produces a mean MOS-LQO that is not statistically significantly lower than half-rate MBE coding alone.

For both full-rate and half-rate MBE, the addition of a worst-case VoLTE link (i.e., AMR coding and decoding at 4.75 kbps) causes mean MOS-LQO to drop by about 0.3 units. This can be attributed to the fact that when speech coders are interconnected, speech coding distortions accumulate, and speech quality must drop. In that light, this speech quality reduction can be viewed as an *inherent cost* of communicating across heterogeneous networks.

While the P25 and LTE networks are heterogeneous with respect to speech coding, they are homogeneous in the most fundamental respect: they both carry bits from one location to another. If the speech coding can be made homogenous, the accumulation of speech coding distortions can be eliminated. Next, we discuss this possibility, the benefits, and the associated requirements.

Calls *within* the LTE network should be coded with AMR (or some other speech coder that exploits the broadband nature of LTE to deliver high speech quality). But when a call is required to bridge the P25 and LTE networks, it can be coded only once, using MBE. When MBE coded speech needs to leave the P25 network, rather than processing it with an MBE decoder, it can be passed off as data to the Internet Protocol (IP) Multimedia Subsystem (IMS) of the LTE network. This data would need proper prioritization within the IMS because it represents a real-time service. The required data rate is less than that required by the lowest AMR rate. Additionally, MBE produces an encoded speech frame every 20 ms, which matches the rate at which the IMS normally receives encoded speech frames from the AMR coder. This approach might be described by “Advanced MBE over LTE” (AMBEoLTE), “MBE over LTE,” or “P25 over LTE.”.



(a) Full-rate MBE interconnected with AMR and full-rate MBE alone.

(b) Half-rate MBE interconnected with AMR and half-rate MBE alone.

Figure 3: MOS-LQO and 95 percent confidence intervals for MBE followed by AMR (dark gray) and AMR followed by MBE (light gray).

The LTE terminal would then require an MBE decoder to transform the MBE bitstream into a speech signal. Likewise, the LTE terminal would require an MBE encoder to produce the MBE bitstream that is carried across the LTE network as IMS data and injected into the P25 network. If these requirements were met, then bidirectional communications between LTE and P25 networks could be conducted with a single (MBE) speech coding; there would be *no speech quality penalty* associated with crossing the network boundary.

The AMBEoLTE approach shows even greater advantage when two P25 systems must be connected by an LTE network. The MBE-encoded speech can be passed throughout the IMS of the LTE network from one P25 system to the other. Only a single (MBE) speech coding is required. The alternative would require MBE coding, then AMR coding, then a second MBE coding due to the accumulation of speech coding distortions; the resulting speech quality would be further reduced from what is shown in Figures 3a and 3b. In short, AMBEoLTE can replace three speech coding processes and the associated three layers of distortion with a single speech coding and just a single layer of distortion.

An additional benefit the AMBEoLTE approach offers is that it makes end-to-end encryption a possibility: the MBE bitstream can be encrypted at the encoder and can remain encrypted on its journey across multiple networks until it arrives at the terminal where it is to be decoded.

We also noted that there are many other speech coders that could be considered for use by FirstNet. If higher data rates can be allocated to speech transmission, then higher speech quality and speech intelligibility can be provided. Higher-rate speech coders may also be more robust to the problems [10] that can be caused by the background noises often present in emergency response environments. Higher-rate speech coders can also provide extended audio bandwidth. The AMR speech coder delivers audio extending from 85 Hertz (Hz) up to 2800 to 3600 Hz (the upper limit depends on the AMR mode) [5]. The mode that requires 12.2 kbps covers audio frequencies up to 3600 Hz. The wideband AMR speech coder (AMR-WB) delivers audio extending from 50 Hz up to 5700 to 6600 Hz [11] (again depending on the mode). To cover audio frequencies up to 6600 Hz, a rate of 23.85 kbps is required. Finally, AMR-WB+ can extend that high-frequency limit to 19.2 kilohertz (kHz), requiring 36.0 kbps. AMR-WB+ stereo coding is also available and requires 48 kbps [12] and [13]. Another family of speech coders is specified in ITU-T Recommendation G.711.1. Here, the audio band from 50 to 4000 Hz is delivered at 64 kbps and the band from 50 to 7000 Hz is delivered at 80 or 96 kbps [14].

We emphasize again that the work presented here provides an initial view of the speech quality issues associated with interconnections of P25 and VoLTE. The use of PESQ allows us to carry out extensive speech quality experiments in an efficient manner. But PESQ shows some limitations when acoustic background noise is present. We are presently engaged in an associated subjective testing effort using public safety practitioners. In this effort, we are using a modified rhyme test to collect actual speech intelligibility results for a subset of the scenarios presented here. The effort includes the effects of several noise environments relevant to public safety operations [15].

5 References

[1] D.W. Griffin and J. S. Lim, “Multiband excitation vocoder,” IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. 36, no. 8, August 1988.

[2] Telecommunications Industry Association, “Project 25 Vocoder Description,” TIA-102.BABA-1998, May 1998. Reaffirmed December 2008.

[3] Telecommunications Industry Association, “APCO Project 25 Half-Rate Vocoder Addendum,” TIA-102.BABA, Annex A, April 2009.

[4] ETSI/3GPP, “AMR speech codec; General description, version 10.0.0,” ETSI TS 26.071, Sophia Antipolis Cedex, France, April 2011.

[5] ETSI, “Performance characterization of the adaptive multi-rate (amr) speech codec,” Tecnical Report TR 126 975, January 2009.

[6] A. Rix, M. Hollier, A. Hekstra, and J. Beerends, “Perceptual evaluation of speech quality (PESQ) – The new ITU standard for end-to-end speech quality assessment, Part I – Time-delay compensation,” J. Audio Engineering Society, vol. 50, no. 10, pp. 755–764, October 2002.

[7] J. Beerends, A. Hekstra, A. Rix, and M. Hollier, “Perceptual evaluation of speech quality (PESQ) – The new ITU standard for end-to-end speech quality assessment, Part II – Psychoacoustic model,” J. Audio Engineering Society, vol. 50, no. 10, pp. 765–778, October 2002.

[8] ITU-T Recommendation P.862, “Perceptual evaluation of speech quality (PESQ), an objective method for end-to end speech quality assessment of narrowband telephone networks and speech codecs,” Geneva, Switzerland, 2001.

[9] IEEE, “IEEE recommended practice for speech quality measurements,” IEEE Transactions on Audio and Electroacoustics, vol. 17, no. 3, pp. 225–246, Sep 1969.

[10] D. Atkinson and A. Catellier, “Intelligibility of selected radio systems in the presence of fireground noise: Test plan and results,” NTIA Technical Report TR-08-453, June 2008. Available at www.its.bldrdoc.gov.

[11] ETSI, “Performance characterization of the adaptive multi-rate wideband (AMR-WB) speech codec,” Technical Report TR 126.976, January 2009.

[12] ETSI/3GPP, “Extended AMR wideband codec; Transcoding functions,” ETSI TS 26.290, Sophia Antipolis Cedex, France, March 2011.

[13] R. Salami, R. Lefebvre, A. Lakaniemi, K. Kontola, S. Bruhn, and A. Taleb, “Extended AMR-WB for high-quality audio on mobile devices,” IEEE Communications Magazine, vol. 44, no. 5, pp.90-97, May 2006.

[14] ITU-T Recommendation G.711.1, “Wideband embedded extension for ITU-T G.711 pulse code modulation,” Geneva, Switzerland, 2008.

[15] NTIA Technical Report TR 13-493: Intelligibility of the Adaptive Multi-Rate Speech Coder in Emergency-Response Environments, December 2012, David J. Atkinson, Stephen D. Voran, and Andrew A. Catellier (http://www.its.bldrdoc.gov/publications/2693.aspx)